Available at www.ComputerScienceWeb.com
POWERED BY SCIENCE DIRECT®

Theoretical Computer Science 302 (2003) 431–456

Theoretical Computer Science

www.elsevier.com/locate/tcs

On the number of occurrences of a symbol in words of regular languages☆

Alberto Bertoni[a], Christian Choffrut[b], Massimiliano Goldwurm[a,*], Violetta Lonati[a]

[a] *Dipartimento di Scienze dell'Informazione, Università degli Studi di Milano, Via Comelico 39, I-20135 Milano, Italy*

[b] *Laboratoire d'Informatique Algorithmique, Fondements et Applications, Université Paris VII, 2 Place Jussieu, 75251 Paris, France*

Received 13 November 2001; received in revised form 8 October 2002; accepted 22 November 2002
Communicated by A. Del Lungo

Abstract

We study the random variable Y_n representing the number of occurrences of a symbol a in a word of length n chosen at random in a regular language $L \subseteq \{a,b\}^*$, where the random choice is defined via a non-negative rational formal series r of support L. Assuming that the transition matrix associated with r is primitive we obtain asymptotic estimates for the mean value and the variance of Y_n and present a central limit theorem for its distribution. Under a further condition on such a matrix, we also derive an asymptotic approximation of the discrete Fourier transform of Y_n that allows to prove a local limit theorem for Y_n. Further consequences of our analysis concern the growth of the coefficients in rational formal series; in particular, it turns out that, for a wide class of regular languages L, the maximum number of words of length n in L having the same number of occurrences of a given symbol is of the order of growth $\lambda^n/\sqrt{n}$, for some constant $\lambda > 1$.

Keywords: Automata and formal languages; Discrete Fourier transform; Gaussian limit distribution; Perron–Frobenius theory; Rational formal series

☆ This work has been supported by the Project M.I.U.R. COFIN "Formal languages and automata: theory and applications".

* Corresponding author. Tel.: +39-02-55006305; fax: +39-02-55006373.

E-mail addresses: bertoni@dsi.unimi.it (A. Bertoni), christian.choffrut@liafa.jussieu.fr (C. Choffrut), goldwurm@dsi.unimi.it (M. Goldwurm), lonati@dsi.unimi.it (V. Lonati).

0304-3975/03/$ - see front matter
doi:10.1016/S0304-3975(02)00892-7

1. Introduction

The aim of this work is to asymptotically estimate the distribution of the number Y_n of occurrences of a given letter in a random text of length n. We consider probability distributions obtained, up to normalization, via $\mathbb{R}_+$-rational series, thus extending the so-called Markovian model [18]. In particular, if the series is the characteristic function of a regular language, then our problem reduces to asymptotically estimate the number of words of length n in a regular language L with k occurrences of the given symbol.

Results concerning this problem are meaningful in three different contexts. The first one is related to the study of ambiguity in formal languages: it is well-known that, for any rational series in non-commutative variables $r \in \mathbb{N}\langle\langle \Sigma \rangle\rangle$, the maximal value of the coefficients associated with words of length n [1] grows as either an exponential or a polynomial expression with respect to n [20,19,16] . This property does not hold in the algebraic case since it has been recently proved that there are context-free grammars that have a logarithmic degree of ambiguity [23]. Similarly, the result cannot be extended to rational series over free partially commutative monoids [8]: from an example given in [23] one can prove that there exists a regular trace language, defined over the monoid generated by the independence alphabet $(\{a,b,c,d\}, \{(a,b),(b,a),(b,c),(c,b),(c,d),(d,c),(a,d),(d,a)\})$, that has a logarithmic ambiguity degree. In our paper we study the asymptotic behaviour of the ambiguity degree of rational series over the free commutative monoid with two generators. This is equivalent to estimating the largest coefficient associated with a monomial of degree n in a rational bivariate function. In particular we show that this value is of the order $\Theta(\lambda^n/\sqrt{n})$, with $\lambda > 1$, for a non-trivial subclass of rational series.

Another context for which our problem is significant concerns the design and the analysis of algorithms for random generation. It is well known that random generation and approximate counting are deeply related [14]. Asymptotic results of the type we present, can be useful to design algorithms for the random generation of words in a regular language containing a given number of occurrences of each letter [7]. We also recall that the degree of ambiguity in a context-free language is a critical parameter for the performance of algorithms for random generation. For instance, a polynomial time algorithm for this problem exists whenever the ambiguity of the language is polynomially bounded. For finitely ambiguous context-free languages, the problem can be solved in an average time of the same order of growth as in the unambiguous case [5].

A third research area involved by our work concerns the analysis of pattern statistics in combinatorics of words (see for instance [12,2,17,18]). There, the main goal is the analysis of the number of occurrences of one or various patterns in a random string generated by a given stochastic process. Many pattern statistic are analysed in the case of random words generated by symmetric Bernoulli [12], Bernoulli or Markov processes [18]. In particular in [17] a pattern statistics is analysed which represents the number of (positions of) occurrences of words from a regular language in a random string of

[1] I.e. the value $\max\{(r,w) \mid w \in \Sigma^*, |w| = n\}$ where, as usual, we denote by (r,w) the coefficients of r associated with the word $w \in \Sigma^*$.

length n generated by a Bernoulli or a Markov process. Many results in this area give conditions under which such statistics asymptotically have a normal distribution in the sense of the central or local limit theorem [11,1]. Our model allows in particular to analyse pattern statistics representing the number of words of a regular language in a random text generated by processes which are slightly more general than Markov processes (see Section 2.1).

We now proceed to state our stochastic model. Let r be a $\mathbb{R}_+$-rational formal series in the non-commutative variables a, b and let n be a positive integer; we denote by Y_n the random variable representing the number of occurrences of a in a word $w \in \{a,b\}^*$ of length n randomly generated with probability $(r,w)/c_r(n)$, where $c_r(n) = \sum_{|x|=n}(r,x)$. E.g., when r is the characteristic series of a language L, w is randomly chosen in $L \cap \{a,b\}^n$ under uniform distribution. Our main aim is to estimate the asymptotic distribution of Y_n, its mean value and variance for the following significant subclass of rational series. Consider a linear representation for the series r, let m be its size and let $\mathscr{A}$ and $\mathscr{B}$ be the $m \times m$ real matrices associated with symbols a and b, respectively. In order to avoid trivial cases we assume that both $\mathscr{A}$ and $\mathscr{B}$ are different from 0. Our main assumption is that the matrix $\mathscr{M} = \mathscr{A} + \mathscr{B}$ is primitive (i.e., there exists an integer k such that all entries of $\mathscr{M}^k$ are non-zero). Then, by the Perron–Frobenius Theorem, $\mathscr{M}$ admits a unique eigenvalue λ of largest modulus. Under the present assumptions, we are able to determine expressions for the mean value and the variance of Y_n of the form

$$\mathbb{E}(Y_n) = \beta n + \mathrm{O}(1), \qquad \mathbb{V}ar(Y_n) = \alpha n + \mathrm{O}(1), \tag{1}$$

where α and β are positive constants. We then use these evaluations to prove a central limit theorem showing that Y_n approximates the normal distribution of mean value βn and variance αn. Under a further mild condition on the matrix $\mathscr{M}$, we also show a local limit theorem for the probability function of Y_n in the sense of the DeMoivre–Laplace Theorem (see for instance [11, Section 12]). More precisely, suppose the following condition holds:

$$|v| < \lambda \text{ for every eigenvalue } v \text{ of } \mathscr{A}\mathrm{e}^{\mathrm{i}\theta} + \mathscr{B} \text{ and every } 0 < \theta < 2\pi. \tag{2}$$

Then, as n tends to $+\infty$, the relation

$$\Pr\{Y_n = k\} = \frac{\mathrm{e}^{-(k-\beta n)^2/2\alpha n}}{\sqrt{2\pi\alpha n}} + \mathrm{o}\left(\frac{1}{\sqrt{n}}\right) \tag{3}$$

holds uniformly for every $k \in \{0,1,\ldots,n\}$.

Even if our problem in its generality does not seem to have been explicitly studied before, several properties we present here already appeared in various forms in the literature and some of our results can be actually obtained by applying more general analyses. In particular, once relations (1) and inequalities $\alpha \neq 0 \neq \beta$ are proved, the central limit theorem for Y_n can be proved by applying a general result due to Bender [1] together with a so-called "quasi-power" theorem whose proof is actually sketched in [17]. Analogously, the local limit property (3) can be obtained as a particular case of Theorem 9.10 in [10] proved by using the machinery of saddle point method [9].

The main contributions of our work are: (i) a direct elementary proof of the variability condition $\alpha>0$, (ii) a precise expression of α and β as functions of $\mathscr{A}$ and of the left and right eigenvectors of $\mathscr{M}$, and (iii) a new proof of the local limit theorem for Y_n based on the use of the discrete Fourier transform and on an approximation of the characteristic function of Y_n. We also stress the fact that condition (2) is necessary to prove (3) in the sense that there exist "primitive" rational series r that do not satisfy (2), for which relation (3) does not hold. We give an example of such a series in the last section and refer back to a companion paper [4] for the study of this phenomenon which is related to a special notion of symbol-periodicity for finite automaton that extends the standard notion of periodicity of non-negative matrices.

At last, we observe that relation (3) has several consequences in the analysis of the ambiguity of formal series. In particular, it implies

$$\max_{0\leqslant k\leqslant n}\left\{\sum_{|w|=n,|w|_a=k}(r,w)\right\}=\Theta\left(\frac{\lambda^n}{\sqrt{n}}\right) \tag{4}$$

and hence, if r is the characteristic series of a language $L\subseteq\{a,b\}^*$ then

$$\max_{0\leqslant k\leqslant n}\{\ \sharp\{x\in L\cap\{a,b\}^n\mid |x|_a=k\}\ \}=\Theta\left(\frac{\lambda^n}{\sqrt{n}}\right). \tag{5}$$

In the context of trace theory, relation (5) yields the growth of the degree of ambiguity of the trace language generated by L over the commutative monoid with generators $\{a,b\}$.

This paper is organized as follows. Section 2 contains a precise statement of the problem which shows in particular that the analysis of the statistics studied in [17] can be reduced to the study of the behaviour of Y_n in some special cases. It recalls the basics of the theory of matrices with positive entries with the Perron–Frobenius theorem and the notion of discrete Fourier transform. Then we study the asymptotic behaviour of the random variable Y_n assuming the matrix M primitive. In Section 3 we first determine the asymptotic expressions for the mean value and the variance. In Section 4 we state the central limit theorem and discuss the main points of its proof. Finally, in Section 5 we give our approximation of the characteristic function of Y_n and prove the local limit theorem for its probability function.

2. Preliminaries

2.1. Statement of the problem

In this section we state our problem formally and study how it compares to those, already considered in the literature, related with pattern occurrence counting. This requires the definition of a probability space given via a so-called rational series, which we now turn to recall (see [3] for more details).

A formal series r over the non-commutative variables a, b with coefficients in the semi-ring of non-negative reals $\mathbb{R}_+$ is a function $r: \{a,b\}^* \to \mathbb{R}_+$, usually represented in the form

$$\sum_{w\in\{a,b\}^*} (r,w)w,$$

where each coefficient (r,w) denotes the value of r at the point w. It is well-known that the set $\mathbb{R}_+\langle\langle a,b\rangle\rangle$ of all such formal series forms a semi-ring w.r.t. the operations of sum and Cauchy product. We recall that r is rational if for some positive integer m there exists a monoid morphism $\mu: \{a,b\}^* \to \mathbb{R}_+^{m\times m}$ and a pair of vectors ξ, η of size m with entries in $\mathbb{R}_+$ such that $(r,w) = \xi'\mu(w)\eta$ for every $w\in\{a,b\}^+$. We say that the triple (ξ,μ,η) is a *linear representation* of r. If the components of ξ, $\mu(a)$, $\mu(b)$ and η are in $\{0,1\}$ then a linear representation can be viewed as a non-deterministic finite automaton.

We now present our model formally. Let r be a $\mathbb{R}_+$-rational formal series in the non-commutative variables a, b and consider a positive $n\in\mathbb{N}$ such that $(r,w)\neq 0$ for some string $w\in\{a,b\}^*$ of length n. Then, for every integer $0\leqslant k\leqslant n$, set

$$\varphi_k^{(n)} = \sum_{|w|=n,|w|_a=k} (r,w) \tag{6}$$

and define the random variable (r.v.) Y_n such that

$$\Pr\{Y_n = k\} = \frac{\varphi_k^{(n)}}{\sum_{j=0}^{n}\varphi_j^{(n)}}. \tag{7}$$

The *rational symbol frequency* problem (r.s.f.p.) consists in studying the distribution of the r.v. Y_n associated with the rational series r.

Example 1. Consider the following representation:

$$\xi' = (10), \quad \mu(a) = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, \quad \mu(b) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad \eta = \begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

The rational series r thus defined satisfies

$$(r,w) = \begin{cases} 1 & \text{if } |w|_a \leqslant 1 \\ 0 & \text{otherwise} \end{cases}$$

leading to the probability distribution

$$\Pr\{Y_n = k\} = \begin{cases} \frac{1}{n+1} & \text{if } k = 0 \\ \frac{n}{n+1} & \text{if } k = 1 \\ 0 & \text{otherwise.} \end{cases}$$

In order to compare the present problem with those previously dealt with in the literature, we show how our model can be viewed as a proper extension of the

Markovian model as far as counting the occurrences of a regular set in a random text is concerned. We refer to [18,17] for a more complete exposition of the Markov models, in the case of a single and of various patterns, respectively. Recall that a Markov process over an alphabet Σ, is used to produce a random sequence of letters. Formally, in the present setting, it is given by a pair (π, P), where $P = \{p_{\sigma,\tau}\}_{\sigma,\tau\in\Sigma\times\Sigma}$ is a stochastic matrix and π is a stochastic vector indexed by Σ. The matrix is interpreted as saying that if the letter at position k is σ, then with probability $p_{\sigma,\tau}$ the letter at position $k+1$ is τ and the value π_σ as the probability of the first letter to be equal to σ. The pair (π, P) induce a probability distribution Π_n over Σ^n

$$\Pi_n(x_1 \dots x_n) = \pi_{x_1} p_{x_1,x_2} \dots p_{x_{n-1},x_n}.$$

Now we are given a regular set of patterns $R \subseteq \Sigma^*$ and we are asked to count the number $O_n(x_1 \dots x_n)$ of occurrences of R in a random text $x_1 \dots x_n$ generated by the above Markov process, where by occurrence is meant a position k in the text where a match with an element of R ends. Observe that the values of $O_n(x_1 \dots x_n)$ range from 0 to n.

In this context the *Markovian pattern frequency problem* consists in studying the distribution

$$\Pr\{O_n = k\}$$

associated with the triple (π, P, R).

Now we show how this problem can be translated into the problem we tackle in this paper. Given the triple (π, P, R), we construct a triple (ξ, μ, η) representing a rational series r such that the r.v. Y_n associated with r satisfies the following equality for every $k = 0, 1, \dots, n$:

$$\Pr\{O_n = k\} = \Pr\{Y_n = k\}.$$

We first construct a (fully defined) finite deterministic automaton recognizing $\Sigma^* R$ whose set of states is Q, the initial state is p and set of final states is F. As usual, we denote by $q \cdot \sigma$ the transition defined by the letter σ in state $q \in Q$. Define the linear representation $\mu : \{a, b\}^* \to \mathbb{R}_+^{Q' \times Q'}$ where $Q' = \{p\} \cup \{(q, \sigma) \mid q \in Q, \sigma \in \Sigma\}$ and all entries of the matrices $\mu(a)$ and $\mu(b)$ are zero except the entries of the form

$$\mu(x)_{p,(q',\sigma)} = \pi_\sigma \quad \text{and} \quad \mu(x)_{(q,\sigma),(q'\tau)} = p_{\sigma,\tau}$$

such that $p \cdot \sigma = q'$ and $q \cdot \tau = q'$ respectively, and (in both cases)

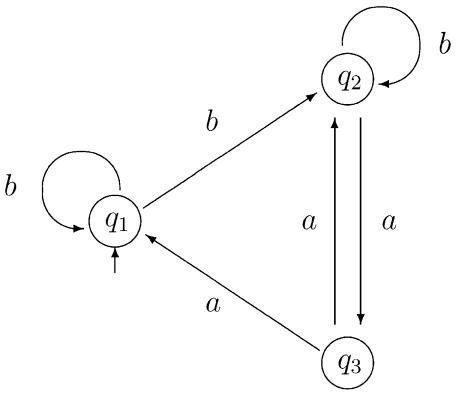
$$x = \begin{cases} a, & \text{if } q' \in F, \\ b, & \text{otherwise.} \end{cases}$$

Denoting by ξ and $\mathbf{1}$ the characteristic vector of $\{p\}$ and Q' respectively, the triple we were looking for is $(\xi, \mu, \mathbf{1})$. Indeed, it is easy to see that

$$\Pr\{Y_n = k\} = \frac{\sum_{w\in\{a,b\}^n, |w|_a=k} \xi\mu(w)\mathbf{1}}{\sum_{w\in\{a,b\}^n} \xi\mu(w)\mathbf{1}} = \frac{\sum_{w\in\Sigma^n, |w|_R=k} \Pi_n(w)}{\sum_{w\in\Sigma^n} \Pi_n(w)} = \Pr\{O_n = k\}.$$

In the case of a Bernoulli model a simplified construction is described by the following example.

Example 2. Let $\Sigma = \{a, b\}$, define $R = \{ba, baa, bab\}$, $\pi = (\frac{1}{2}, \frac{1}{2})$ and let $P = [p_{\sigma\tau}]$ be given by $p_{\sigma\tau} = \frac{1}{2}$ for every $\sigma, \tau \in \Sigma$. Consider the Markovian pattern frequency problem associated with (π, P, R). To define an equivalent *r.s.f.p.* let T be the smallest deterministic finite automaton recognizing the language $\Sigma^* R$. Let us transform T so that each transition entering a final state is labelled by a and any other transition is labelled by b. Consider as final all states of the new automaton and reduce it by collapsing into a unique state any pair of equivalent states.[2] We obtain the following (weighted) non-deterministic finite automaton where all states are final and all transitions have weight $\frac{1}{2}$.



The corresponding linear representation (ξ, μ, η) defines a *r.s.f.p.* equivalent to the original problem.

The converse does not hold in general. Indeed, for each triple (π, P, R) carrying out the previous construction yields a rational series in non-commutative variables a, b. Taking its commutative image

$$\sum_{n=0}^{\infty} \sum_{k=0}^{n} \Pr\{O_n = k\} x^k y^n$$

and considering it as a real valued function in two variables, it can be easily verified that it is rational in the variables x, y. On the other hand, consider the rational series

[2] I.e., those with equal outgoing transitions.

of Example 1. It leads to the non-rational function

$$\sum_{n=0}^{\infty}\sum_{k=0}^{n}\Pr\{Y_n=k\}x^k y^n=(x-1)\frac{\log(1-y)}{y}+\frac{x}{1-y}$$

which cannot be obtained in the Markovian model.

Now, we introduce the formalism we use throughout this paper. Let $\mathscr{A}$ and $\mathscr{B}$ be matrices associated with the symbols a and b, respectively, i.e., $\mathscr{A}=\mu(a)$ and $\mathscr{B}=\mu(b)$. To avoid trivial cases we assume that both $\mathscr{A}$ and $\mathscr{B}$ are non-null. Then, the commutative image of r is given by the bivariate series

$$\sum_{n=0}^{\infty}\sum_{k=0}^{n}\varphi_k^{(n)}a^k b^{n-k}=\sum_{n=0}^{\infty}\xi'(\mathscr{A}a+\mathscr{B}b)^n\eta.$$

Replacing the variables a and b by the monomials xy and y, respectively, we obtain the series

$$\sum_{n=0}^{\infty}\left(\sum_{k=0}^{n}\varphi_k^{(n)}x^k\right)y^n=\sum_{n=0}^{\infty}\xi'(\mathscr{A}x+\mathscr{B})^n\eta\cdot y^n \tag{8}$$

which implies that the values $\varphi_k^{(n)}$ are defined by

$$\sum_{k=0}^{n}\varphi_k^{(n)}x^k=\xi'(\mathscr{A}x+\mathscr{B})^n\eta. \tag{9}$$

In the next two subsection present two classical mathematical tools we use throughout this work. The first one is the Perron–Frobenius Theory for non-negative matrices while the second is the discrete Fourier transform.

2.2. The Perron–Frobenius theorem

The Perron–Frobenius theory is a well-known subject widely studied in the literature (see for instance [21]). To recall its main results we first fix some notation. For every pair of matrices T,S, the expression $T>S$ means that $T_{ij}>S_{ij}$ for every pair of indices i,j. As usual, we consider any vector v as a column vector and denote by v' the corresponding row vector. We recall that a non-negative matrix T is called *primitive* if there exists $m\in\mathbb{N}$ such that $T^m>0$. The main properties of such matrices are given by the following theorem [21, Section 1].

Theorem 1 (Perron–Frobenius). *Let T be a primitive non-negative matrix. There exists an eigenvalue λ of T (called Perron–Frobenius eigenvalue of T) such that*:

(1) *λ is real and positive*;
(2) *with λ we can associate strictly positive left and right eigenvector*;
(3) *$|v|<\lambda$ for every eigenvalue $v\neq\lambda$*;
(4) *if $0\leqslant C\leqslant T$ and γ is an eigenvalue of C, then $|\gamma|\leqslant\lambda$; moreover $|\gamma|=\lambda$ implies $C=T$*;
(5) *λ is a simple root of the characteristic polynomial of T.*

The following proposition is a direct application of the theorem above [21, Exercise 1.9].

Proposition 1. *Let $C=\{c_{ij}\}$ be a $m\times m$ complex matrix and let $T=\{t_{ij}\}$ be a primitive matrix of the same size, such that $|c_{ij}|\leqslant t_{ij}$ for every i,j. If λ is the Perron–Frobenius eigenvalue of T, then for every eigenvalue γ of C we have $|\gamma|\leqslant\lambda$. Moreover, if $|\gamma|=\lambda$ for some eigenvalue γ of C, then $|c_{ij}|=t_{ij}$ for every i,j.*

Another consequence of the Perron–Frobenius Theorem concerns the asymptotic growth of the entries of the nth power of a primitive matrix T; this is of the order $\Theta(\lambda^n)$, where λ is the Perron–Frobenius eigenvalue of T. More precisely, the following property holds [21, Theorem 1.2].

Proposition 2. *If T is a primitive matrix and 1 is its Perron–Frobenius eigenvalue, then*

$$T^n = uv' + C(n)\cdot\frac{n^s}{h^n} + \mathrm{o}\left(\frac{n^s}{h^n}\right) \quad \text{for } n\to+\infty,$$

where $s\in\mathbb{N}$, $h>1$, $C(n)$ is a complex matrix such that $|C(n)_{ij}|\leqslant c$ (for a fixed constant c and for any i,j,n) and v' and u are strictly positive left and right eigenvectors of T corresponding to the eigenvalue 1, normed so that $v'u=1$.

Throughout this work we assume the matrix $\mathscr{M}=\mathscr{A}+\mathscr{B}$ primitive. Then, by the Perron–Frobenius Theorem, $\mathscr{M}$ admits exactly one eigenvalue λ of maximum modulus, which is real and positive. Dividing $\mathscr{M},\mathscr{A},\mathscr{B}$ and $\varphi_k^{(n)}$ by λ, we get

$$M=\frac{\mathscr{M}}{\lambda},\quad A=\frac{\mathscr{A}}{\lambda},\quad B=\frac{\mathscr{B}}{\lambda},\qquad \phi_k^{(n)}=\frac{\varphi_k^{(n)}}{\lambda^n}. \tag{10}$$

In particular, Eqs. (7) and (9) become

$$\Pr\{Y_n=k\}=\frac{\phi_k^{(n)}}{\sum_{j=0}^{n}\phi_j^{(n)}} \tag{11}$$

and

$$\sum_{k=0}^{n}\phi_k^{(n)}x^k=\xi'(Ax+B)^n\eta. \tag{12}$$

2.3. The Fourier transform

The characteristic function of a random variable X is defined by

$$F_X(\theta)=\mathbb{E}(\mathrm{e}^{\mathrm{i}\theta X}).$$

Thus, if X is a discrete r.v. assuming values in $\mathbb{N}$, then

$$F_X(\theta)=\sum_{k\in\mathbb{N}}\Pr\{X=k\}\,\mathrm{e}^{\mathrm{i}\theta k}.$$

We recall that F_X is always well-defined for every $\theta\in\mathbb{R}$, it is periodic of period 2π and it completely characterizes the r.v. X. Moreover it represents the classical tool to prove convergence in distribution: a sequence of r.v.'s $\{X_n\}_n$ converges in distribution to a r.v. X (i.e. $\lim_{n\to\infty}\Pr\{X_n\leqslant t\}=\Pr\{X\leqslant t\}$ for every $t\in\mathbb{R}$) if and only if $F_{X_n}(\theta)$ tends to $F_X(\theta)$ for every $\theta\in\mathbb{R}$. Several forms of the central limit theorem are classically proved in this way [11].

In the following, for the sake of brevity, we denote by $F_n(\theta)$ the characteristic function of Y_n:

$$F_n(\theta)=\frac{\sum_{k=0}^{n}\phi_k^{(n)}e^{i\theta k}}{\sum_{j=0}^{n}\phi_j^{(n)}}. \tag{13}$$

Now, let us recall the definition of the discrete Fourier transform (for more details see [6]). For any positive integer n, the nth discrete Fourier transform is the transformation $D_n:\mathbb{C}^n\to\mathbb{C}^n$ such that, for every $u=(u_0,\ldots,u_{n-1})\in\mathbb{C}^n$, $D_n(u)=(v_0,\ldots,v_{n-1})$ where

$$(D_n(u))_s=v_s=\sum_{k=0}^{n-1}\omega_n^{sk}u_k,$$

ω_n being the nth principal root of unity (i.e. $\omega_n=e^{i2\pi/n}$). It is well-known that D_n admits an inverse transformation D_n^{-1} given by

$$(D_n^{-1}(v))_k=\frac{1}{n}\sum_{s=0}^{n-1}\omega_n^{-ks}v_s.$$

We may apply these notions to the coefficients $\phi_k^{(n)}$ defined in (12): consider the vector $(\phi_0^{(n)},\ldots,\phi_n^{(n)})$ as an element of $\mathbb{C}^{n+1}$ and let

$$(f_0^{(n)},\ldots,f_n^{(n)})=D_{n+1}(\phi_0^{(n)},\ldots,\phi_n^{(n)})$$

be its discrete Fourier transform. Then

$$f_s^{(n)}=\sum_{k=0}^{n}\omega_{n+1}^{sk}\phi_k^{(n)}\quad\text{for } s=0,\ldots,n.$$

Thus, when applying D^{-1} to the vector $(f_0^{(n)},\ldots,f_n^{(n)})\in\mathbb{C}^{n+1}$, we have

$$\phi_k^{(n)}=\frac{1}{n+1}\sum_{s=0}^{n}\omega_{n+1}^{-ks}f_s^{(n)}\quad\text{for } k=0,\ldots,n. \tag{14}$$

In this way we obtain an explicit expression for the values $\phi_k^{(n)}$, while in (12) they just appear as coefficients of a polynomial. Observe that, for each $s=0,\ldots,n$, the component $f_s^{(n)}$ satisfies the relation

$$f_s^{(n)}=\left(\sum_{j=0}^{n}\phi_j^{(n)}\right)\cdot F_n\left(\frac{2\pi s}{n+1}\right) \tag{15}$$

and can be directly computed from the matrices A and B:

$$f_s^{(n)} = \sum_{k=0}^{n} \omega_{n+1}^{sk} \phi_k^{(n)} = \xi'(A e^{i2\pi s/n+1} + B)^n \eta. \tag{16}$$

We will use these coefficients $f_s^{(n)}$ in Section 5 in order to prove a local limit theorem for the probability distribution of Y_n. This is given by an asymptotic evaluation of the coefficients $\phi_k^{(n)}$ obtained by first determining an approximation of the values $F_n(2\pi s/n+1)$ in (15) and then applying (14).

3. Analysis of mean value and variance

In this section we give an asymptotic evaluation of the expected value and of the variance of Y_n. To this end consider the function

$$h_n(z) = \sum_{k=0}^{n} \phi_k^{(n)} e^{zk} = \xi'(A e^z + B)^n \eta. \tag{17}$$

Note that $h_n(z)$ is related to the characteristic function of Y_n:

$$F_n(\theta) = \frac{h_n(i\theta)}{h_n(0)}.$$

It is also well-known that the first two moments of Y_n can be obtained by evaluating h_n and its derivatives at $z=0$:

$$\mathbb{E}(Y_n) = \frac{h_n'(0)}{h_n(0)}, \quad \mathbb{E}(Y_n^2) = \frac{h_n''(0)}{h_n(0)}. \tag{18}$$

From (17), recalling the algebra of matrices is non-commutative, one easily obtains the following equations:

$$h_n(0) = \xi' M^n \eta, \quad h_n'(0) = \xi' \sum_{i=0}^{n-1} M^i A M^{n-1-i} \eta, \tag{19}$$

$$h_n''(0) - h_n'(0) = 2\xi' \sum_{l=0}^{n-2} \sum_{r=0}^{n-2-l} M^l A M^r A M^{n-2-r-l} \eta. \tag{20}$$

Now, since M is a primitive matrix with maximum eigenvalue 1, by Proposition 2 we have

$$M^n = uv' + C(n) \cdot \frac{n^s}{h^n}, \tag{21}$$

where $s \in \mathbb{N}$, $h > 1$, $C(n)$ is a complex matrix such that $|C(n)_{ij}| \leqslant c$ (for a fixed constant c and for any i, j, n) and v' and u are strictly positive left and right eigenvectors of M corresponding to the eigenvalue 1, normed so that $v'u = 1$. Moreover, the following matrix is well defined:

$$C = \sum_{n=0}^{\infty} C(n) \frac{n^s}{h^n}. \tag{22}$$

Thus, replacing (21) in Eqs. (19) and (20), a rather long but conceptually simple computation shows the following

Lemma 1.

$$h_n(0) = (\xi' u)(v'\eta) + \mathrm{O}\left(\frac{n^s}{h^n}\right),$$

$$h_n'(0) = n(\xi' u)(v'Au)(v'\eta) + (\xi' CAu)(v'\eta) + (\xi' u)(v'AC\eta) + \mathrm{O}\left(\frac{n^{2s+1}}{h^n}\right),$$

$$\begin{aligned} h_n''(0) = {} & n(n-1)(\xi' u)(v'Au)^2(v'\eta) + n(\xi' u)(v'Au)(v'\eta) \\ & + 2n[(\xi' CAu)(v'Au)(v'\eta) + (\xi' u)(v'ACAu)(v'\eta) + (\xi' u)(v'Au)(v'AC\eta)] \\ & + \mathrm{O}(1). \end{aligned}$$

This allows us to evaluate the mean value and the variance of Y_n as function of the matrices A and C and of the eigenvectors v' and u of M. The following theorem is easily derived from (18) by applying the previous lemma.

Theorem 2. *There exist $s \in \mathbb{N}$ and $h > 1$ such that*

$$\mathbb{E}(Y_n) = (v'Au)n + \frac{\xi' CAu}{\xi' u} + \frac{v'AC\eta}{v'\eta} + \mathrm{O}\left(\frac{n^{2s+1}}{h^n}\right), \tag{23}$$

$$\mathbb{V}ar(Y_n) = \{(v'Au) - (v'Au)^2 + 2(v'ACAu)\}n + \mathrm{O}(1). \tag{24}$$

Observe that in the case where we count the number of occurrences of the letter a in a random word belonging to a rational language L, then the main terms of both $\mathbb{E}(Y_n)$ and $\mathbb{V}ar(Y_n)$ do not depend on the initial and final states of the automaton recognizing L. If, further, the automaton is totally defined[3] then $\mathbb{E}(Y_n) = n/2 + \mathrm{O}(1)$.

Moreover, both $\mathbb{E}(Y_n)$ and $\mathbb{V}ar(Y_n)$ always have strictly linear behaviour as shown now.

Theorem 3. *The constants of the main terms of mean value and variance of Y_n*

$$\beta = v'Au, \quad \alpha = (v'Au) - (v'Au)^2 + 2(v'ACAu) \tag{25}$$

are non-null.

Proof. Since A is non-null and both v and u are strictly positive (point 2 of Theorem 1), it is clear that $\beta > 0$. Proving that α is strictly positive is equivalent to proving that $\mathbb{V}ar(Y_n) \geqslant cn$ for some $c > 0$ and for infinitely many n. Recall that the r.v. Y_n is associated with the polynomial $\xi'(Ax + B)^n\eta$ having coefficients in $\mathbb{R}_+$ and degree

[3] I.e. for every state q and every $\sigma \in \{a, b\}$ there is exactly one state reachable from q by an arrow labelled by σ.

equal to n. This leads us to consider for any non-null polynomial $p(x)=\sum_{k\in I} p_k x^k$, where $I\subseteq\mathbb{N}$ and $p_k\geqslant 0$ for each $k\in I$, the associated random variable X_p such that $\Pr\{X_p=k\}=p_k/p(1)$. Let $V(p)$ be the variance of X_p. Then

$$V(p)=\frac{p''(1)+p'(1)}{p(1)}-\left(\frac{p'(1)}{p(1)}\right)^2. \tag{26}$$

Claim. *For any pair of non-null polynomials p,q with non-negative coefficients, we have*

$$V(pq)=V(p)+V(q),\quad V(p+q)\geqslant\frac{p(1)}{p(1)+q(1)}V(p)+\frac{q(1)}{p(1)+q(1)}V(q).$$

In particular, we have $V(p+q)\geqslant\min\{V(p),V(q)\}$.

Proof. The first equation follows immediately from (26). Further, observe that

$$(p(1)+q(1))V(p+q)=p''(1)+q''(1)+p'(1)+q'(1)-\frac{(p'(1)+q'(1))^2}{p(1)+q(1)}.$$

Thus, the second relation follows again from (26) by recalling that $(a+b)^2/c+d\leqslant a^2/c+b^2/d$, for every four-tuple of positive values a,b,c,d. □

Now, let us return to the Theorem. Since $A+B$ is primitive and both A and B are non-null, there exists an integer t such that all the entries of the matrix $C=(Ax+B)^t$ are polynomials with at least two non-null coefficients. This implies that the value

$$c=\min\{V(C_{ij})\,|\,i,j=1,2,\ldots,m\}$$

is strictly positive. Then, by the previous claim, for every $n\in\mathbb{N}$ and every pair of indices i,j we have

$$V(C_{ij}^{n+1})\geqslant\min\{V(C_{ik}^n)+V(C_{kj})\,|\,k=1,2,\ldots,m\}.$$

As a consequence, $V(C_{ij}^{n+1})\geqslant c+\min\{V(C_{ik}^n)\,|\,k=1,2,\ldots,m\}$ proving that $V(C_{ij}^n)\geqslant nc$. Since $\xi'(Ax+B)^n\eta$ is a polynomial associated with the r.v. Y_n, we get

$$\mathbb{V}ar(Y_{tn})\geqslant\min\{V(C_{ij}^n)\,|\,i,j=1,2,\ldots,m\}\geqslant nc$$

for every $n\in\mathbb{N}$. Together with (24) this proves $\mathbb{V}ar(Y_n)=\Theta(n)$ and hence $\alpha>0$. □

4. Integral limit theorem

In this section we show that a central limit theorem for the sequence $\{Y_n\}$ holds under the simple hypothesis that M is primitive. The result can be obtained by applying Theorem 1 in [1] and is actually a special case of a more general analysis [10]. However we present it here in some details for sake of completeness and because several identities encountered in the proof will be useful in the following section. In

particular we will need some properties used to prove Proposition 3 and its corollary; that proposition is a quasi-power theorem of the type studied in [13] and its proof, sketched in [17], applies to our case also.

First we study the behaviour of $h_n(z)$ near $z=0$; to this end consider the bivariate generating function of $\{\phi_k^{(n)}\}_{k,n}$, given by

$$G(x,y)=\sum_{n=0}^{+\infty}\sum_{k=0}^{n}\phi_k^{(n)}x^k y^n=\sum_{n=0}^{+\infty}\xi'(Ax+B)^n\eta\cdot y^n.$$

Such a function can be written in the form

$$G(x,y)=\xi'(I-y(Ax+B))^{-1}\eta=\frac{Q(x,y)}{P(x,y)},$$

where $P(x,y)=\mathrm{Det}(I-y(Ax+B))$ and $Q(x,y)=\xi'\,\mathrm{Adj}(I-y(Ax+B))\eta$. Moreover, observe that $G(e^z,y)$ is also the generating function of the sequence $\{h_n(z)\}_n$; hence we have

$$G(e^z,y)=\sum_{n=0}^{+\infty}h_n(z)y^n=\frac{Q(e^z,y)}{P(e^z,y)}.$$

Now recall that $A+B$ is primitive and 1 is its eigenvalue of largest modulus; then, by the Perron–Frobenius Theorem, the equation

$$\mathrm{Det}(uI-(A\mathrm{e}^z+B))=0 \tag{27}$$

implicitly defines an analytic function $u=u(z)$ in a neighbourhood of $z=0$ such that $u(0)=1$ and $u'(0)\neq 0$. Such a function satisfies the following property:

Proposition 3. *There are two positive constants c, ρ and a function $R(z)$ non-null at $z=0$, rational with respect to e^z and $u(z)$, such that for every $|z|\leqslant c$*

$$h_n(z)=R(z)u(z)^n+\mathrm{O}(\rho^n)$$

and $\rho<|u(z)|$.

Proof. By the Perron–Frobenius Theorem and a continuity property, there exists $\rho>0$ such that, for every z near 0, all roots μ of (27) different from 1 (i.e. all other eigenvalues of $A\mathrm{e}^z+B$) satisfy the relation $|\mu|<\rho<|u(z)|$.

This means that, for a suitable constant $c>0$ and for every $|z|\leqslant c$, the polynomial (w.r.t. the variable y)

$$\frac{\mathrm{Det}(I-y(A\mathrm{e}^z+B))}{1-u(z)y}=\frac{P(e^z,y)}{1-u(z)y}$$

only has roots of modulus greater than $\rho^{-1}>|u(z)|^{-1}$.

Thus, since $Q(1,1)\neq 0$, the function $G(e^z,y)=Q(e^z,y)/P(e^z,y)$ can be expressed in the form

$$G(e^z,y)=\frac{R(z)}{1-u(z)y}+E(z,y), \tag{28}$$

where, for every $|z|\leqslant c$, $E(z,y)$ has singularities μ^{-1} of modulus greater than ρ^{-1} and, by l'Hôpital's rule, $R(z)$ is given by

$$R(z)=\frac{-u(z)\cdot Q(e^z,u(z)^{-1})}{P'_y(e^z,u(z)^{-1})}. \tag{29}$$

Since $G(e^z,y)$ is the generating function of $\{h_n(z)\}$, from (28) we have

$$h_n(z)=R(z)u(z)^n+E_n(z), \tag{30}$$

where $E_n(z)$ is the nth coefficient of $E(z,y)$. Now, to evaluate $E_n(z)$, let $\mathscr{C}$ be the circle of center in 0 and radius ρ^{-1} and apply Cauchy's integral formula choosing $\mathscr{C}$ as contour; we obtain

$$|E_n(z)|=\left|\frac{1}{2\pi\mathrm{i}}\int_{\mathscr{C}}\frac{E(z,y)}{y^{n+1}}\,\mathrm{d}y\right|\leqslant K(z)\rho^n$$

for a suitable constant $K(z)$. Then, denoting by K the value $\max_{|z|\leqslant c}|K(z)|$, we obtain $|E_n(z)|\leqslant K\rho^n$ and this completes the proof. □

To study the limit distribution of Y_n it is convenient to express its moments as function of $u(z)$ and the corresponding derivatives. We can do that by simply developing the previous proof.

Corollary 1. *For some $0<\rho<1$, we have*

$$\mathbb{E}(Y_n)=u'(0)n+\frac{R'(0)}{R(0)}+\mathrm{O}(\rho^n),$$

$$\mathbb{V}ar(Y_n)=(u''(0)-u'(0)^2)n+\frac{R''(0)}{R(0)}-\left(\frac{R'(0)}{R(0)}\right)^2+\mathrm{O}(\rho^n).$$

Proof. Let $R(z)$ and $u(z)$ be defined as in Proposition 3. Then, for some $0<\rho<1$, we have

$$\begin{aligned}
h_n(0)&=R(0)+\mathrm{O}(\rho^n),\\
h'_n(0)&=R(0)u'(0)n+R'(0)+\mathrm{O}(\rho^n),\\
h''_n(0)&=R(0)u'(0)^2n^2+2\{R'(0)u'(0)+R(0)(u''(0)-u'(0)^2)\}n\\
&\quad+R''(0)+\mathrm{O}(\rho^n).
\end{aligned} \tag{31}$$

Indeed, since $u(0)=1$, the first equation is a straightforward consequence of Proposition 3. Moreover from (30) we get

$$h'_n(z)=(R'(z)u(z)+nR(z)u'(z))u(z)^{n-1}+E'_n(z).$$

Note that $E'(0,y)$ has the same singularities of $E(0,y)$ and hence $E'_n(0)=\mathrm{O}(\rho^n)$ for some $0<\rho<1$. This proves the second equation and the third one follows from a similar reasoning. Thus, the result is a consequence of equations (18). □

By Theorem 3, this corollary implies

$$u'(0)=\beta, \quad u''(0)-u'(0)^2=\alpha. \tag{32}$$

Hence we are able to prove the following

Theorem 4. *If the matrix M is primitive then there exists two positive algebraic numbers α, β such that the r.v. $Y_n-\beta n/\sqrt{\alpha n}$ converges in distribution to the normal r.v. of mean 0 and variance 1, i.e. for every $x\in\mathbb{R}$*

$$\lim_{n\to+\infty}\Pr\left\{\frac{Y_n-\beta n}{\sqrt{\alpha n}}\leqslant x\right\}=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x}\mathrm{e}^{-t^2/2}\,\mathrm{d}t.$$

Proof. We argue as in [1]. Let $\bar{F}_n(t)$ be the characteristic function of $Y_n-\beta n/\sqrt{\alpha n}$. Then

$$\bar{F}_n(t)=\sum_{k=0}^{n}\Pr\{Y_n=k\}\mathrm{e}^{\mathrm{i}t(k-\beta n/\sqrt{\alpha n})}=\frac{\mathrm{e}^{-\mathrm{i}t\beta\sqrt{\frac{n}{\alpha}}}\cdot h_n(\mathrm{i}t/\sqrt{\alpha n})}{h_n(0)}.$$

By Proposition 3, as n tends to $+\infty$, we have (for some $0<\rho<1$)

$$\bar{F}_n(t)=\exp\left\{-\mathrm{i}t\beta\sqrt{\frac{n}{\alpha}}+\log R\left(\frac{\mathrm{i}t}{\sqrt{\alpha n}}\right)+n\log u\left(\frac{\mathrm{i}t}{\sqrt{\alpha n}}\right)-\log R(0)+\mathrm{O}(\rho^n)\right\}. \tag{33}$$

Now observe that the Taylor expansion of $\log u(z)$ at the point $z=0$ is

$$\log u(z)=zu'(0)+\frac{z^2}{2}\{u''(0)-u'(0)^2\}+\mathrm{O}(z^3).$$

Thus, since $R(z)$ is analytic in 0, from (33) and (32) we obtain

$$\bar{F}_n(t)=\exp\left\{-\frac{t^2}{2}+\mathrm{O}\left(\sqrt{\frac{1}{n}}\right)\right\},$$

showing that $\bar{F}_n$ pointwise converges to the characteristic function of the standard normal random variable. □

5. Local limit theorem

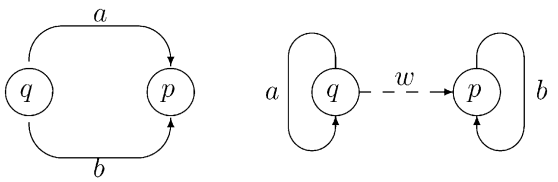
In this section we prove a local limit theorem for the r.v. Y_n, assuming condition (2) stated in the Introduction. Using the notation of Section 2 such a condition can be

stated as follows:

$$|\mu| < 1 \quad \text{for every eigenvalue } \mu \text{ of } A\mathrm{e}^{\mathrm{i}\theta} + B \text{ and every } 0 < \theta < 2\pi. \tag{34}$$

Note that such hypothesis is often verified, as the following example shows.

Example 3. Let us consider the formal series having a linear representation defined by a finite automaton over the alphabet $\{a, b\}$ with set of states Q. If the associated matrix $\mathscr{M}$ is primitive and one of the following conditions holds for some pair of distinct states $q, p \in Q$:
(1) $q \cdot a = q \cdot b = p$;
(2) $q \cdot a = q$ and $p \cdot b = p$,
then $\mathscr{A}\mathrm{e}^{\mathrm{i}\theta} + \mathscr{B}$ satisfies condition (2).



Case 1. Case 2.

Proof. (1) Let us define $\mathscr{M}(\theta) = \mathscr{A}\mathrm{e}^{\mathrm{i}\theta} + \mathscr{B}$. Clearly $|\mathscr{M}(\theta)_{ij}| \leqslant \mathscr{M}_{ij}$ for any i, j, θ. Moreover, since $\mathscr{M}(\theta)_{qp} = \mathrm{e}^{\mathrm{i}\theta} + 1$ and $\mathscr{M}_{qp} = 2$, we have $|\mathscr{M}(\theta)_{qp}| \neq \mathscr{M}_{qp}$ for any $\theta \neq 2k\pi$. Therefore, if v is an eigenvalue of $\mathscr{M}(\theta)$, we have $|v| < \lambda$ by Proposition 1.

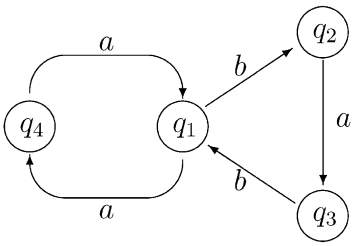
(2) Since $\mathscr{M}$ is primitive, there exists a word w such that $\delta(q, w) = p$. Let $h = |w|_a$ and consider the words aw and wb: they have the same length $H = |w| + 1$ and $\delta(q, aw) = \delta(q, wb) = p$. Now we can suppose that case 1 holds for no pair of states (else condition (2) is already proved); then we can compute the polynomial

$$\mathscr{M}(\theta)^H_{qp} = \sum_{|v|=H,\ \delta(q,v)=p} \mathrm{e}^{\mathrm{i}\theta|v|_a} = \mathrm{e}^{\mathrm{i}\theta h} + \mathrm{e}^{\mathrm{i}\theta(h+1)} + \cdots = \mathrm{e}^{\mathrm{i}\theta h}(1 + \mathrm{e}^{\mathrm{i}\theta}) + \cdots.$$

Then, since $\mathscr{M}^H_{qp} \in \mathbb{N}$, we have $|\mathscr{M}(\theta)^H_{qp}| \neq \mathscr{M}^H_{qp}$ for any $\theta \neq 2k\pi$, while clearly $|\mathscr{M}(\theta)^H_{ij}| \leqslant \mathscr{M}^H_{ij}$ for any i, j, θ. Now, since $\mathscr{M}$ is primitive and λ is its Perron–Frobenius eigenvalue, $\mathscr{M}^H$ is also primitive and λ^H is its Perron–Frobenius eigenvalue. On the other hand, if v is an eigenvalue of $\mathscr{M}(\theta)$ for some $\theta \neq 2k\pi$, then also v^H is an eigenvalue of $\mathscr{M}(\theta)^H$; therefore $|v^H| < \lambda^H$ by Proposition 1 and this proves condition (2). □

We now present an example where condition (2) does not hold.

Example 4. Consider the following automaton



and the associated varied transition matrix $\mathscr{M}(\theta)=\mathscr{A}e^{i\theta}+\mathscr{B}$. Then the matrices $\mathscr{M}(\pi/2)$, $\mathscr{M}(\pi)$, $\mathscr{M}(3\pi/2)$ admit eigenvalues of modulus equal to λ.

Proof. Indeed the matrix $\mathscr{M}(\theta)$ and its characteristic polynomial are given by

$$\mathscr{M}(\theta)=\begin{pmatrix} 0 & 1 & 0 & e^{i\theta} \\ 0 & 0 & e^{i\theta} & 0 \\ 1 & 0 & 0 & 0 \\ e^{i\theta} & 0 & 0 & 0 \end{pmatrix}, \quad \mathrm{Det}(yI-\mathscr{M}(\theta))=y^4-y^2e^{i2\theta}-ye^{i\theta}.$$

Let λ be the Perron–Frobenius eigenvalue of $\mathscr{M}$, hence $\lambda^4-\lambda^2-\lambda=0$. Thus, it is easy to prove that $-i\lambda$ is a root of the polynomial $\mathrm{Det}(yI-\mathscr{M}(\pi/2))=y^4+y^2-iy$. Analogously, $-\lambda$ and $i\lambda$ are roots of the polynomials $\mathrm{Det}(yI-\mathscr{M}(\pi))$ and $\mathrm{Det}(yI-\mathscr{M}(3\pi/2))$, respectively. □

We now illustrate the proof of the local limit theorem for Y_n. This consists of three main steps. First we study the behaviour of the function $\lambda(\theta)^n$ near 0, $\lambda(\theta)$ being the eigenvalue of maximum modulus of $Ae^{i\theta}+B$. Then, in Section 5.2 we use this analysis to obtain a pointwise approximation of the function F_n, which in view of (15) applies to the coefficients $f_s^{(n)}$ as well. This result will lead in Section 5.3 to an evaluation of the values $\phi_k^{(n)}$ by anti-transforming the coefficients $f_s^{(n)}$ via Eq. (14).

5.1. Main eigenvalue analysis

In this subsection we study the behaviour of the function

$$\lambda(\theta)=u(i\theta)$$

in a real neighbourhood of $\theta=0$, where u is implicitly defined by Equation (27). For this reason, we do not need here the hypothesis (2). Also observe that, by Proposition 3, $\lambda(\theta)$ determines the main term of the characteristic function of Y_n.

First note that, by a continuity property, $\lambda(\theta)$ is the eigenvalue of maximum modulus of $Ae^{i\theta}+B$ for any θ near 0. Also, for every non-null θ near 0, $Ae^{i\theta}+B\neq M$ and hence $|\lambda(\theta)|<1$ by Proposition 1. Therefore, we can state the following.

Proposition 4. *There exists $\theta_0>0$ such that, for every $|\theta|<\theta_0$, $Ae^{i\theta}+B$ has a unique eigenvalue of maximum modulus $\lambda(\theta)$; further, in that neighbourhood, $\lambda(\theta)$ is an analytic function and $|\lambda(\theta)|<1$ for every $0<|\theta|<\theta_0$.*

In the next statement the constant θ_0 is the same as in the previous Proposition.

Proposition 5. *The function $|\lambda(\theta)|$ is even while $\arg\lambda(\theta)$ is odd, i.e. $|\lambda(\theta)|=|\lambda(-\theta)|$ and $\arg\lambda(-\theta)=-\arg\lambda(\theta)$. Moreover, for every $|\theta|<\theta_0$,*

$$|\lambda(\theta)| = 1-\frac{\alpha}{2}\theta^2+O(\theta^4) \quad \text{and} \quad \arg\lambda(\theta)=\beta\theta+O(\theta^3), \tag{35}$$

where α and β are defined in (25).

Proof. Observe that the polynomial $D(x,y)=\mathrm{Det}(yI-(Ax+B))$ has real coefficients; then, denoting by $\bar{z}$ the conjugate of z, we have that $D(x,y)=0$ implies $D(\bar{x},\bar{y})=0$. Therefore, by (27), $\overline{\lambda(\theta)}=u(-i\theta)=\lambda(-\theta)$ and hence

$$\mathrm{Re}\,\lambda(-\theta)=\tfrac{1}{2}(\lambda(-\theta)+\lambda(\theta))=\mathrm{Re}\,\lambda(\theta),$$

$$\mathrm{Im}\,\lambda(-\theta)=\frac{1}{2i}(\lambda(-\theta)-\lambda(\theta))=-\mathrm{Im}\,\lambda(\theta).$$

As a consequence we obtain

$$\arg\lambda(-\theta)=\mathrm{arctg}\,\frac{\mathrm{Im}\,\lambda(-\theta)}{\mathrm{Re}\,\lambda(-\theta)}=-\arg\lambda(\theta),$$

$$|\lambda(-\theta)|^2=(\mathrm{Re}\,\lambda(-\theta))^2+(\mathrm{Im}\,\lambda(-\theta))^2=|\lambda(\theta)|^2.$$

Hence, there exist two constants $a,b\in\mathbb{R}$ such that

$$|\lambda(\theta)| = 1-\frac{a}{2}\theta^2+O(\theta^4) \quad \text{and} \quad \arg\lambda(\theta)=b\theta+O(\theta^3). \tag{36}$$

However, by the definition of $\lambda(\theta)$ and by (32) we have $\lambda'(0)=iu'(0)=i\beta$, while by (36)

$$\lambda'(0)=\left[\frac{\partial}{\partial\theta}(|\lambda(\theta)|e^{i\arg\lambda(\theta)})\right]_{\theta=0}=ib$$

and hence $b=\beta$. Analogously, by (32) we have $\lambda''(0)=-u''(0)=-\alpha-\beta^2$ while, computing $\lambda''(0)$ from (36), we get $\lambda''(0)=-a-b^2$ yielding $a=\alpha$. □

Proposition 6. *For any $\frac{1}{3}<\varepsilon<\frac{1}{2}$ we have*

$$|\lambda^n(\theta)-e^{-(\alpha/2)\theta^2 n+i\beta\theta n}|=O\left(\frac{1}{\sqrt{n}}\right) \quad \textit{for every } |\theta|\leqslant\frac{2\pi}{(n+1)^{\varepsilon}}.$$

Proof. The previous proposition implies $\lambda(\theta) \sim e^{-(\alpha/2)\theta^2+i\beta\theta+ic\theta^3}$ as $\theta \to 0$, for some constant c. Then, for any $\varepsilon > \frac{1}{3}$ and for all $|\theta| \leqslant 2\pi/(n+1)^{\varepsilon}$,

$$\lambda^n(\theta) \sim e^{-(\alpha/2)\theta^2 n+i\beta\theta n}(1+ic\theta^3 n).$$

Hence

$$|\lambda^n(\theta) - e^{-(\alpha/2)\theta^2 n+i\beta\theta n}| = O(n|\theta|^3 e^{-(\alpha/2)\theta^2 n}) = O\left(\frac{1}{\sqrt{n}}\right),$$

where the last inequality is obtained by deriving the middle term with respect to θ. □

5.2. Approximating the Fourier transform

In this section we study the characteristic function of Y_n and in particular the term

$$h_n(i\theta) = \xi'(Ae^{i\theta} + B)^n \eta.$$

Proposition 7. *For every $0 < \theta_0 < \pi$ there exists $0 < \tau < 1$ such that*

$$h_n(i\theta) = O(\tau^n) \quad n \to +\infty$$

for all $\theta_0 \leqslant |\theta| \leqslant \pi$.

Proof. As shown in Section 4, the generating function of $\{h_n(i\theta)\}$ is given by

$$G(e^{i\theta}, y) = \sum_{n=0}^{\infty} h_n(i\theta) y^n = \frac{Q(e^{i\theta}, y)}{P(e^{i\theta}, y)}.$$

Observe that the singularities of $G(e^{i\theta}, y)$ are the roots of $P(e^{i\theta}, y)$, i.e. the values μ^{-1}, for each eigenvalue μ of $Ae^{i\theta} + B$. Now, consider an arbitrary $0 < \theta_0 < \pi$. From (34) we know that there exists $0 < \tau < 1$ such that $|\mu| < \tau < 1$ for every $|\theta| \in [\theta_0, \pi]$. Hence, in this interval all singularities of $G(e^{i\theta}, y)$ are in modulus greater than τ^{-1}. Thus, reasoning as in the proof of Proposition 3, the result follows applying Cauchy's integral formula to the function $G(e^{i\theta}, y)$ with integral contour given by $|y| = \tau^{-1}$. □

The previous proposition, together with Proposition 3 and the analysis of function $\lambda(\theta)$ given in the previous subsection, allows us to establish the following theorem, which though technical, is of significant importance for further results.

Theorem 5. *Let θ_0 be the constant defined in Proposition* 4. *Then, for every $\theta \in [-\pi, \pi]$,*

$$|F_n(\theta) - e^{-(\alpha/2)\theta^2 n+i\beta n\theta}| = \Delta_n(\theta),$$

where

$$\Delta_n(\theta) = \begin{cases} \mathrm{O}\left(\frac{1}{n^\varepsilon}\right) & \text{if } |\theta| \leqslant \frac{2\pi}{(n+1)^\varepsilon} \\ \mathrm{O}(\mathrm{e}^{-\alpha\pi^2 n^{1-2\varepsilon}}) & \text{if } \frac{2\pi}{(n+1)^\varepsilon} \leqslant |\theta| \leqslant \theta_0 \\ \mathrm{O}(\tau^n) & \text{if } \theta_0 \leqslant |\theta| \leqslant \pi \end{cases}$$

for some $0<\tau<1$ *and for every* $\frac{1}{3}<\varepsilon<\frac{1}{2}$.

This theorem suggests to approximate $F_n(\theta)$ by the function $\hat{F}_n(\theta)$ defined by:
(i) $\hat{F}_n(\theta)$ is periodic on $\mathbb{R}$ with period 2π;
(ii) $\hat{F}_n(\theta) = \mathrm{e}^{-(\alpha/2)\theta^2 n + \mathrm{i}\beta n\theta}$ for $|\theta| \leqslant \pi$.

Proof. We study separately the three cases

$$|\theta| \leqslant \frac{2\pi}{(n+1)^\varepsilon}, \quad \frac{2\pi}{(n+1)^\varepsilon} \leqslant |\theta| \leqslant \theta_0 \quad \text{and} \quad \theta_0 \leqslant |\theta| \leqslant \pi.$$

Since $F_n(\theta) = h_n(\mathrm{i}\theta)/h_n(0)$, in the first interval by Proposition 3 we have

$$F_n(\theta) = \frac{R(\mathrm{i}\theta)\lambda^n(\theta) + \mathrm{O}(\rho^n)}{R(0) + \mathrm{O}(\rho^n)}.$$

Also observe that $R(\mathrm{i}\theta) = R(0) + \mathrm{O}(n^{-\varepsilon})$ and, by Proposition 6, $\lambda^n(\theta) = \mathrm{e}^{-(\alpha/2)\theta^2 n + \mathrm{i}\beta\theta n} + \mathrm{O}(n^{-1/2})$. Thus, replacing these values in the previous numerator, we get

$$F_n(\theta) = \mathrm{e}^{-(\alpha/2)n\theta^2 + \mathrm{i}\beta n\theta} + \mathrm{O}\left(\frac{1}{n^\varepsilon}\right).$$

Now consider the second interval $2\pi/(n+1)^\varepsilon \leqslant |\theta| \leqslant \theta_0$. By Proposition 3, we have

$$\begin{aligned} &|h_n(\mathrm{i}\theta) - R(0)\mathrm{e}^{-(\alpha/2)n\theta^2 + \mathrm{i}\beta n\theta}| \\ &\quad \leqslant |h_n(\mathrm{i}\theta) - R(\mathrm{i}\theta)\lambda(\theta)^n| + |R(\mathrm{i}\theta)\lambda(\theta)^n - R(0)\mathrm{e}^{-(\alpha/2)n\theta^2 + \mathrm{i}\beta n\theta}| \\ &\quad \leqslant \mathrm{O}(\rho^n) + |R(\mathrm{i}\theta)\lambda(\theta)^n| + R(0)\mathrm{e}^{-(\alpha/2)n\theta^2}. \end{aligned}$$

From Eq. (35) we know that, for some real constant c and for every $|\theta| \leqslant \theta_0$,

$$|\lambda(\theta)| \leqslant 1 - \frac{\alpha}{2}\theta^2 + c\theta^4$$

and this implies $|\lambda(\theta)|^n \leqslant (1 - \frac{\alpha\theta^2}{4})^n \leqslant \mathrm{e}^{-n(\alpha\theta^2/4)}$ for every $|\theta| \leqslant \sqrt{\alpha/4c}$. Since we may assume that $\theta_0 \leqslant \sqrt{\alpha/4c}$, the inequality above yields

$$|h_n(\mathrm{i}\theta) - R(0)\mathrm{e}^{-(\alpha/2)n\theta^2 + \mathrm{i}\beta n\theta}| \leqslant \mathrm{O}(\rho^n + \mathrm{e}^{-n(\alpha\theta^2/4)}) \leqslant \mathrm{O}(\mathrm{e}^{-\pi^2\alpha n^{1-2\varepsilon}}).$$

Finally, assume $\theta_0 \leqslant |\theta| \leqslant \pi$. In this case, by Proposition 7, we obtain

$$|h_n(\mathrm{i}\theta) - R(0)\mathrm{e}^{-(\alpha/2)n\theta^2+\mathrm{i}\beta n\theta}| \leqslant |h_n(\mathrm{i}\theta)| + |R(0)|\mathrm{e}^{-(\alpha/2)n\theta^2} = \mathrm{O}(\tau^n + \mathrm{e}^{-(\alpha/2)\theta_0^2 n})$$

which proves the result. □

Now, Eqs. (17) and (31) imply

$$\sum_{k=0}^{n} \phi_k^{(n)} \sim R(0) \tag{37}$$

as n tends to $+\infty$. Hence, applying the previous result to Eq. (15), we get

Corollary 2. *For every integer s, $0 \leqslant s \leqslant n$, we have*

$$\left| f_s^{(n)} - R(0)\hat{F}_n\left(\frac{2\pi s}{n+1}\right)\right| = \begin{cases} \Delta_n\left(\frac{2\pi s}{n+1}\right) & \text{if } 0 \leqslant s \leqslant \frac{n+1}{2} \\ \Delta_n\left(\frac{2\pi s}{n+1} - 2\pi\right) & \text{if } \frac{n+1}{2} < s \leqslant n \end{cases}$$

where Δ_n is defined as in Theorem 5.

5.3. Anti-transform

In this section we estimate the coefficients $\phi_k^{(n)}$ by using Eq. (14) and applying the results of the previous subsection. In particular, we use Corollary 2 to approximate the coefficients $f_s^{(n)}$.

First, let us define the array $(\hat{\phi}_0^{(n)}, \ldots, \hat{\phi}_n^{(n)})$ as the inverse discrete Fourier transform of the vector of components

$$R(0)\hat{F}_n\left(\frac{2\pi s}{n+1}\right) \quad \text{for } s = 0, \ldots, n.$$

Then, by classical tools of analysis, we get for every $k = 0, \ldots, n$

$$\begin{aligned}
\hat{\phi}_k^{(n)} &= \frac{R(0)}{n+1} \sum_{s=0}^{n} \hat{F}_n\left(\frac{2\pi s}{n+1}\right) \cdot \omega_{n+1}^{-ks} \\
&\approx R(0) \int_0^1 \hat{F}_n(2\pi u)\, \mathrm{e}^{-\mathrm{i}2\pi k u}\, \mathrm{d}u \\
&= R(0) \int_{-\frac{1}{2}}^{\frac{1}{2}} \hat{F}_n(2\pi u)\, \mathrm{e}^{-\mathrm{i}2\pi k u}\, \mathrm{d}u \\
&= R(0) \int_{-\frac{1}{2}}^{\frac{1}{2}} \mathrm{e}^{-2\alpha\pi^2 n u^2} \mathrm{e}^{-\mathrm{i}2\pi(k-\beta n)u}\, \mathrm{d}u \\
&\approx R(0) \int_{-\infty}^{+\infty} \mathrm{e}^{-2\alpha\pi^2 n u^2} \mathrm{e}^{-\mathrm{i}2\pi(k-\beta n)u}\, \mathrm{d}u.
\end{aligned}$$

Recalling that

$$\int_{-\infty}^{+\infty} \mathrm{e}^{-\delta u^2} \cdot \mathrm{e}^{-\mathrm{i}2\pi m u}\,\mathrm{d}u = \sqrt{\frac{\pi}{\delta}}\mathrm{e}^{-(\pi^2/\delta)m^2},$$

the last expression can be reduced to

$$\hat{\phi}_k^{(n)} = \frac{R(0)}{\sqrt{2\alpha n\pi}}\mathrm{e}^{-(k-\beta n)^2/2\alpha n} + \mathrm{o}\left(\frac{1}{\sqrt{n}}\right). \tag{38}$$

This gives an approximation of our coefficients $\phi_k^{(n)}$ and the following proposition shows the associated error bound, which does not depend on k.

Proposition 8. *There exists a positive constant c such that, for every n large enough, the following relation holds uniformly for every $k=0,1,\ldots,n$ and for any $\frac{1}{3}<\varepsilon<\frac{1}{2}$:*

$$|\phi_k^{(n)} - \hat{\phi}_k^{(n)}| \leqslant \frac{c}{n^{2\varepsilon}}.$$

Proof. By the definition of anti-transform and by Corollary 2 we have

$$\begin{aligned}|\phi_k^{(n)} - \hat{\phi}_k^{(n)}| &\leqslant \frac{1}{n+1}\sum_{s=0}^{n}\left|f_s^{(n)} - R(0)\hat{F}_n\left(\frac{2\pi s}{n+1}\right)\right| \\ &\leqslant \frac{2}{n+1}\sum_{0\leqslant s\leqslant n+1/2} \Delta_n\left(\frac{2\pi s}{n+1}\right) + \mathrm{O}(\rho^n)\end{aligned}$$

for some $0<\rho<1$. The last sum can be computed by splitting the range of s in three parts, corresponding to the three intervals studied in Theorem 5. We obtain

$$\begin{aligned}&|\phi_k^{(n)} - \hat{\phi}_k^{(n)}| \\ &\quad= \mathrm{O}\left(\frac{1}{n}\left[n^{1-\varepsilon}n^{-\varepsilon} + \left(\frac{\theta_0}{2\pi} - n^{-\varepsilon}\right)n\,\mathrm{e}^{-\alpha\pi^2 n^{1-2\varepsilon}} + \left(\frac{1}{2} - \frac{\theta_0}{2\pi}\right)n\tau^n\right]\right) \\ &\quad= \mathrm{O}(n^{-2\varepsilon} + \mathrm{e}^{-\alpha\pi^2 n^{1-2\varepsilon}} + \tau^n) = \mathrm{O}\left(\frac{1}{n^{2\varepsilon}}\right). \qquad \square\end{aligned}$$

To conclude, we can summarize our main result in the following theorem.

Theorem 6. *Assume that the matrix $\mathscr{M}$ is primitive, the matrices $\mathscr{A}$ and $\mathscr{B}$ are different from 0 and $|v|<\lambda$ for every eigenvalue v of $\mathscr{A}\mathrm{e}^{\mathrm{i}\theta}+\mathscr{B}$ and every $0<\theta<2\pi$. Then, there exist two positive constants α and β such that, as n tends to $+\infty$,*

the equation

$$\Pr\{Y_n = k\} = \frac{e^{-(k-\beta n)^2/2\alpha n}}{\sqrt{2\pi\alpha n}} + o\left(\frac{1}{\sqrt{n}}\right), \tag{39}$$

holds uniformly for every $k = 0, 1, \ldots, n$.

Proof. By Theorem 3, the constants α and β are strictly positive. Then, equality (39) follows from (11) and (37) by applying Proposition 8 and relation (38). □

Finally, from the previous theorem one can easily deduce the following.

Corollary 3. *Let* $r \in \mathbb{R}_+\langle\langle a, b\rangle\rangle$ *be a rational formal series that satisfies the hypothesis of Theorem* 6. *Then, as* n *tends to* $+\infty$, *the largest value* $\sum_{|w|=n,\,|w|_a=k}(r, w)$ *for* $0 \leqslant k \leqslant n$ *is of the order of growth* $\Theta(\lambda^n/\sqrt{n})$, *for some* $\lambda > 1$.

6. Conclusions

In this work we have studied the rational symbol frequency problem for formal series defined via a primitive matrix. In particular, we have proved a local limit theorem for the associated r.v. Y_n stating that, if condition (2) is satisfied, then the probability function of Y_n approximates a normal density function. Intuitively this result means that, in our hypotheses, the occurrence of the letter a in a given position of a "random" word of length n is rather independent of the other occurrences and of the position itself. Thus, Y_n is similar to the sum of n independent Bernoulli random variables of equal parameter.

We observe that the primitivity hypothesis cannot be omitted to obtain a Gaussian limit distribution. To see this fact, it is sufficient to consider the language a^*b^*.

Also condition (2) is necessary to obtain the local limit result in the sense that if it does not hold Eq. (39) may not be true. As an example, consider the r.f.s.p. defined by the weighted automaton of Example 2. As shown there, this is equivalent to a Markovian pattern frequency problem where the pattern is $R = \{ba, baa, bab\}$ and the stochastic model is given by a Bernoulli process of parameter $\frac{1}{2}$. In this case the characteristic polynomial of the matrix $\mathscr{A} + \mathscr{B}$ is $y^2(y-1)$ and hence its Perron–Frobenius eigenvalue is 1 (implying $\mathscr{A} = A$ and $\mathscr{B} = B$). Moreover, the characteristic polynomial of $\mathscr{A}e^{i\theta} + \mathscr{B}$ is

$$\mathrm{Det}(Iy - \mathscr{A}e^{i\theta} - \mathscr{B}) = y\left(y^2 - y + \frac{1 - e^{2i\theta}}{4}\right)$$

the roots of which are 0, $1 + e^{i\theta}/2$ and $1 - e^{i\theta}/2$. Thus 1 is eigenvalue of the matrix also for $\theta = \pi$ showing that condition (2) is not true in this case.

On the other hand, the probability function of the associated r.v. Y_n does not satisfy relation (39). Indeed, in this case the bivariate generating function of $\{\phi_k^{(n)}\}$ is

given by

$$G(x,y)=\xi'(I-y(\mathscr{A}x+\mathscr{B}))^{-1}\eta=\frac{xy^2-x^2y^2+4}{y^2-x^2y^2-4y+4}.$$

Hence, we can directly compute its coefficients obtaining (for $2\leqslant k<n$)

$$\Pr\{Y_n=k\}=\frac{1}{2^n}\begin{cases}\binom{n}{k}+\binom{n}{k+1} & \\ \quad -\binom{n-2}{k-2}-\binom{n-2}{k-1} & \text{if } k \text{ is even}\\ \binom{n-2}{k-1}+\binom{n-2}{k} & \text{if } k \text{ is odd and } k<n-1\end{cases}$$

which clearly cannot approximate a Gaussian density (even if, the automaton being primitive, a central limit theorem holds for Y_n).

The local limit distribution of Y_n depends on structure properties of the automaton described in Example 2 we investigate in a forthcoming paper [4].

References

[1] E.A. Bender, Central and local limit theorems applied to asymptotic enumeration, J. Combin. Theory 15 (1973) 91–111.

[2] E.A. Bender, F. Kochman, The distribution of subword counts is usually normal, European J. Combin. 14 (1993) 265–275.

[3] J. Berstel, C. Reutenauer, Rational Series and their Languages, Springer, New York, Heidelberg, Berlin, 1988.

[4] A. Bertoni, C. Choffrut, M. Goldwurm, V. Lonati, The symbol-periodicity of irreducible finite automata, Rapporto Interno no. 277-02, Dipartimento di Scienze dell'Informazione, Università degli Studi di Milano, April 2002 (available at http://homes.dsi.unimi.it/∼goldwurm/inglese.html).

[5] A. Bertoni, M. Goldwurm, M. Santini, Random generation for finitely ambiguous context-free languages, RAIRO: Theoretical Informatics and Applications, Vol. 35, 2001, pp. 499–512.

[6] R.N. Bracewell, The Fourier Transform and its Applications, McGraw-Hill Book Company, New York, 1986.

[7] A. Denise, O. Roques, M. Termier, Random generation of words of context-free languages according to the frequencies of letters, Trends in Mathematics, Birkauser, 2000.

[8] V. Diekert, G. Rozenberg (Eds.), The Book of Traces, World Scientific, Singapore, 1995.

[9] P. Flajolet, R. Sedgewick, The average case analysis of algorithms: saddle point asymptotics, Rapport de recherche no. 2376, INRIA Rocquencourt, October 1994.

[10] P. Flajolet, R. Sedgewick, The average case analysis of algorithms: multivariate asymptotics and limit distributions, Rapport de recherche no. 3162, INRIA Rocquencourt, May 1997.

[11] B.V. Gnedenko, The Theory of Probability (translated by G. Yankovsky), Mir Publishers, Moscow, 1976.

[12] L.J. Guibas, A.M. Odlyzko, String overlaps, pattern matching, and nontransitive games, J. Combin. Theory Ser. A 30 (2) (1981) 183–208.

[13] H.K. Hwang, Théorèmes limites pour les structures combinatoires et les fonctions arithmétiques, Ph.D. Dissertation, École Polytechnique, Palaiseau, France, 1994.

[14] M.R. Jerrum, L.G. Valiant, V.V. Vazirani, Random generation of combinatorial structures from a uniform distribution, Theoret. Comput. Sci. 43 (2–3) (1986) 169–188.
[15] J. Kemeny, J.L. Snell, Finite Markov Chains, D. Van Nostrand Co., Inc, Princeton, Toronto, New York, 1960.
[16] W. Kuich, G. Baron, Two papers on automata theory, Technical Report no. 253, Institute fur Informationsverarbeitung, Technische Universität Graz und Österreich Computer Gesellschaft, June 1988.
[17] P. Nicodeme, B. Salvy, P. Flajolet, Motif statistics, in: J. Nešetřil (Ed.), Proc. 7th ESA, Lecture Notes in Computer Science, Vol. 1643, Springer, Berlin, 1999, pp. 194–211.
[18] M. Régnier, W. Szpankowski, On pattern frequency occurrences in a Markovian sequence, Algorithmica 22 (4) (1998) 621–649.
[19] C. Reutenauer, Propriétés arithmétiques et topologiques de séries rationnelles en variables non commutatives, These Sc. Maths, Doctorat troisieme cycle, Université Paris VI, 1977.
[20] M.P. Schützenberger, Certain elementary family of automata, in: Proc. Symp. Mathematical Theory of Automata, 1962, pp. 139–153.
[21] E. Seneta, Non-negative Matrices and Markov Chains, Springer, New York, Heidelberg, Berlin, 1981.
[22] K. Wich, Exponential ambiguity of context-free grammars, in: G. Rozenberg, W. Thomas (Eds.), Proc. 4th DLT, World Scientific, Singapore, 2000, pp. 125–138.
[23] K. Wich, Sublinear ambiguity, in: M. Nielsen, B. Rovan (Eds.), Proc. 25th MFCS, Lecture Notes in Computer Science, Vol. 1893, Springer, Berlin, 2000, pp. 690–698.